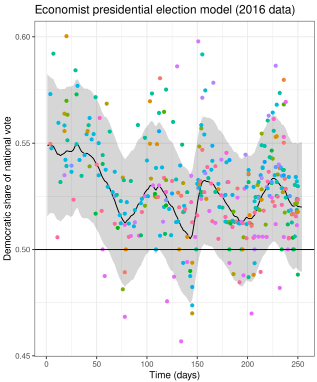
Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

Theory and Foundations of Statistics in the Era of Big Data — Honoring Basu and Bahadur (April 2024)

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

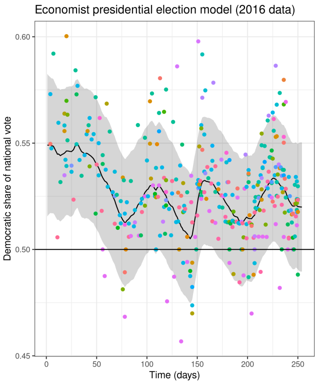
Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

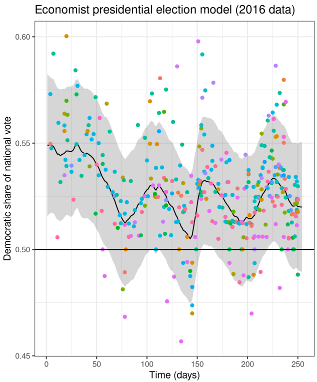
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

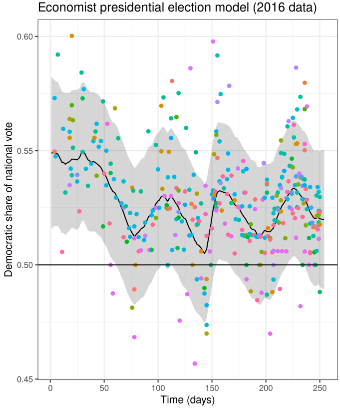
Problem: Each MCMC run takes about 10 hours (Stan, six cores).

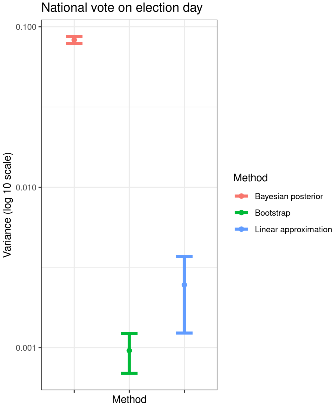
Results

Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

Results

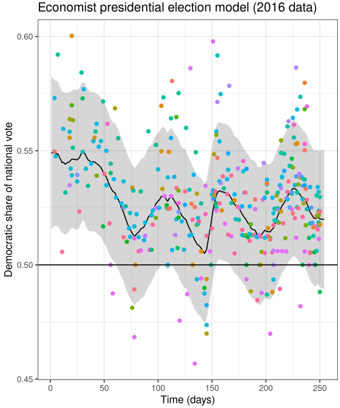
Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

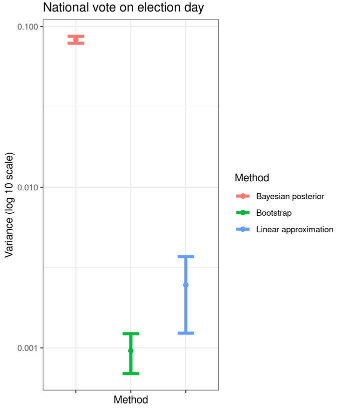




Results

Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.





Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

Outline

- Data reweighting
    - Write the change in the posterior expectation as linear component $+$ error
    - The linear component can be computed from a single run of MCMC

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate, even as $N \to \infty$
- What should the exchangeable unit be?

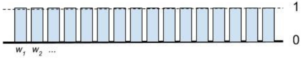
Data re-weighting.

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:

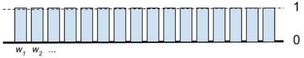


Data re-weighting.

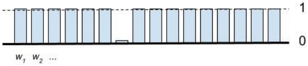
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:

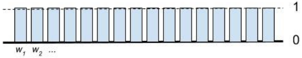


Data re-weighting.

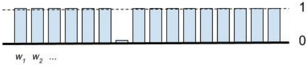
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

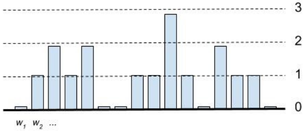
Original weights:



Leave-one-out weights:



Bootstrap weights:

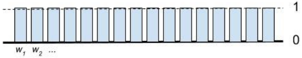


Data re-weighting.

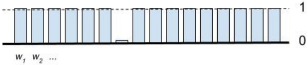
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

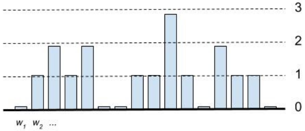
Original weights:

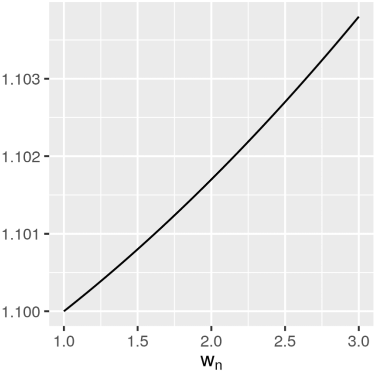


Leave-one-out weights:



Bootstrap weights:



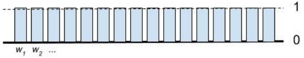


Data re-weighting.

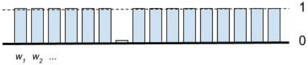
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

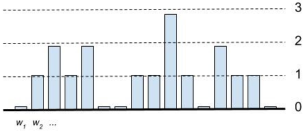
Original weights:

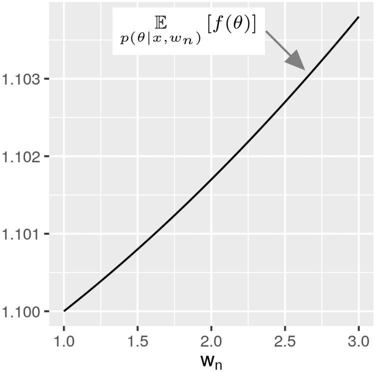


Leave-one-out weights:



Bootstrap weights:



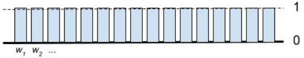


Data re-weighting.

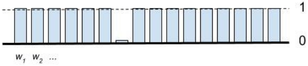
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

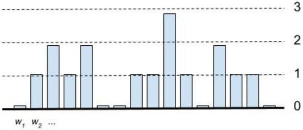
Original weights:

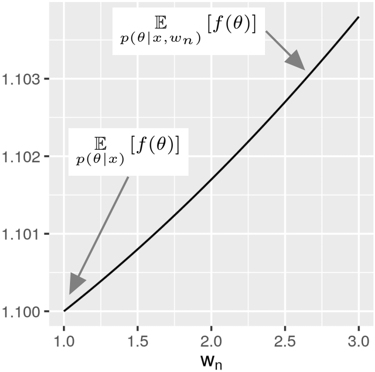


Leave-one-out weights:



Bootstrap weights:

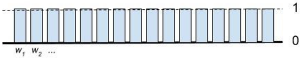




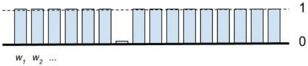
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

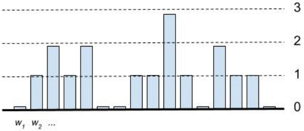
Original weights:

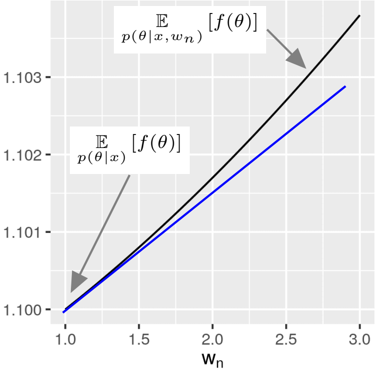


Leave-one-out weights:



Bootstrap weights:



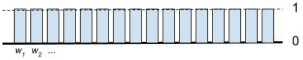


Data re-weighting.

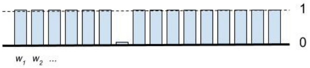
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

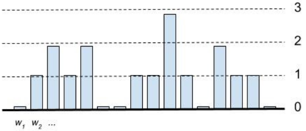
Original weights:

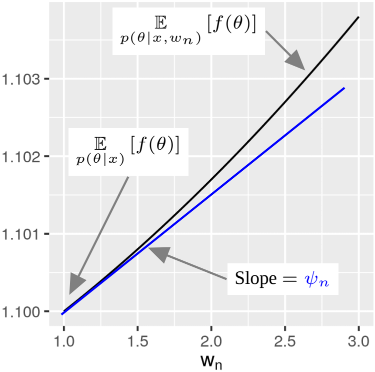


Leave-one-out weights:



Bootstrap weights:





Data re-weighting.

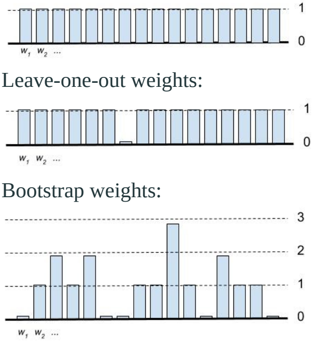
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

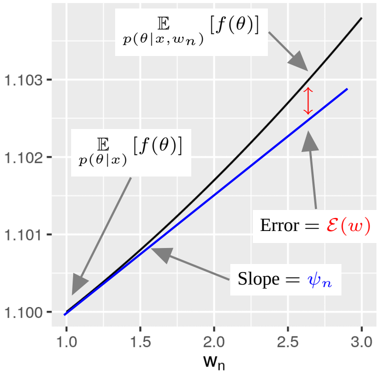
$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:

Leave-one-out weights:

Bootstrap weights:



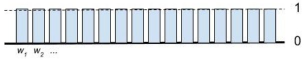


Data re-weighting.

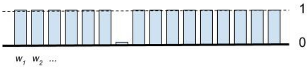
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

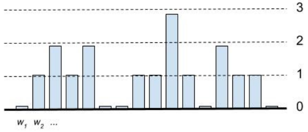
Original weights:

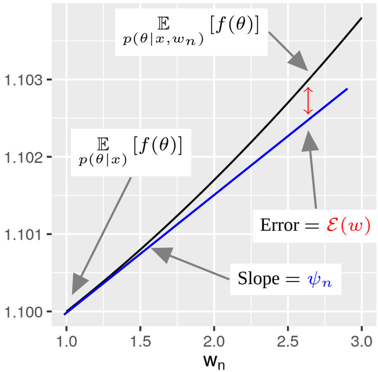


Leave-one-out weights:



Bootstrap weights:





The re-scaled slope $N\psi_n$ is known as the "influence function" at data point x_n.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w)$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathop{\mathbb{E}}_{p(\theta|X,w_n)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathop{\mathbb{E}}_{p(\theta|X,w_n)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\mathop{\mathbb{E}}_{p(\theta|X)}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}(w_n) = \frac{1}{2}\underbrace{\mathop{\mathbb{E}}_{p(\theta|X,\tilde{w}_n)}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}(w_n) = \frac{1}{2}\underbrace{\underset{p(\theta|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

$= O_p(N^{-1})$ under posterior concentration

$= O_p(N^{-2})$ under posterior concentration

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}(w_n) = \frac{1}{2}\underbrace{\underset{p(\theta|X,\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

$= O_p(N^{-1})$ under posterior concentration $\qquad$ $= O_p(N^{-2})$ under posterior concentration

Theorem [Giordano and Broderick, 2023] (paraphrase):
If the posterior $p(\theta|X)$ "concentrates" (e.g. as in the Bernstein–von Mises theorem),[a] then

$$w_n \mapsto N\left(\underset{p(\theta|X,w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]\right)$$

becomes linear as $N \to \infty$, with slope $\lim_{N\to\infty} \psi_n$.

[a] Existing results are sufficient for a *particular weight* [Kass et al., 1990]. Giordano and Broderick [2023] proves that the result holds when averaged over all weights, as needed for variance estimation.

High dimensional problems

What about when parts of the posterior don't concentrate?

Example: **Genearlized linear model with random effects (REs) λ and fixed effect γ.**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

High dimensional problems

What about when parts of the posterior don't concentrate?

Example: **Genearlized linear model with random effects (REs) λ and fixed effect γ.**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

Does $w_n \mapsto \underset{p(\gamma|X,w_n)}{\mathbb{E}}[f(\gamma)] - \underset{p(\gamma|X)}{\mathbb{E}}[f(\gamma)]$ become linear as N grows?

(Note $p(\gamma|X)$ *does* concentrate.)

High dimensional problems

What about when parts of the posterior don't concentrate?

Example: **Genearlized linear model with random effects (REs) λ and fixed effect γ.**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

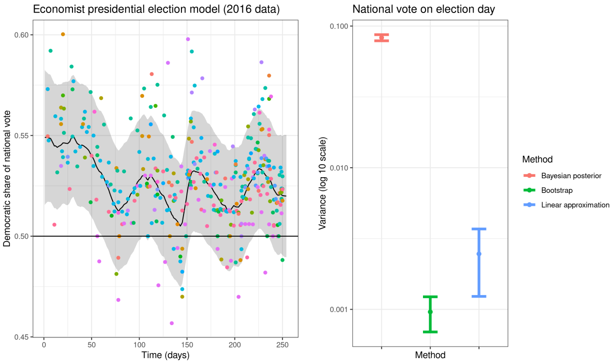
Does $w_n \mapsto \underset{p(\gamma|X,w_n)}{\mathbb{E}}[f(\gamma)] - \underset{p(\gamma|X)}{\mathbb{E}}[f(\gamma)]$ become linear as N grows?
(Note $p(\gamma|X)$ *does* concentrate.)

Theorem 5 of Giordano and Broderick [2023] (paraphrase): In general, **no!**
Specifically, if $p(\lambda|X,\gamma)$ does not concentrate, then
— even if $p(\gamma|X)$ concentrates marginally —
both the slope ψ_n and the error $\mathcal{E}(w_n)$ are $O_p(N^{-1})$, and so

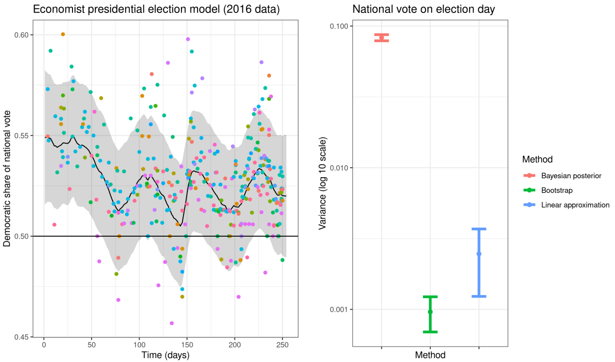
$$N\left(\underset{p(\gamma|X,w_n)}{\mathbb{E}}[f(\gamma)] - \underset{p(\gamma|X)}{\mathbb{E}}[f(\gamma)]\right) = N\psi_n(w_n - 1) + N\mathcal{E}(w_n) \text{ is nonlinear.}$$

However, $\mathcal{E}(w_n) \to 0$ as $\underset{p(\lambda|X,\gamma)}{\mathrm{Cov}}(\lambda) \to 0$.

Observations and consequences



Observations and consequences



- We use often use models of the form $p(\gamma, \lambda | X)$.
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.

  ... Especially given the linear approximation's huge computational advantage.

Preprint: Giordano and Broderick [2023] (`arXiv:2305.06466`)
(The preprint focuses on variance estimation, the present results are found in the proofs.)

A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL `https://projects.economist.com/us-2020-forecast/president`. Data and model accessed Oct., 2020.

R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.

J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.